CTA

computational text analysis from scratch

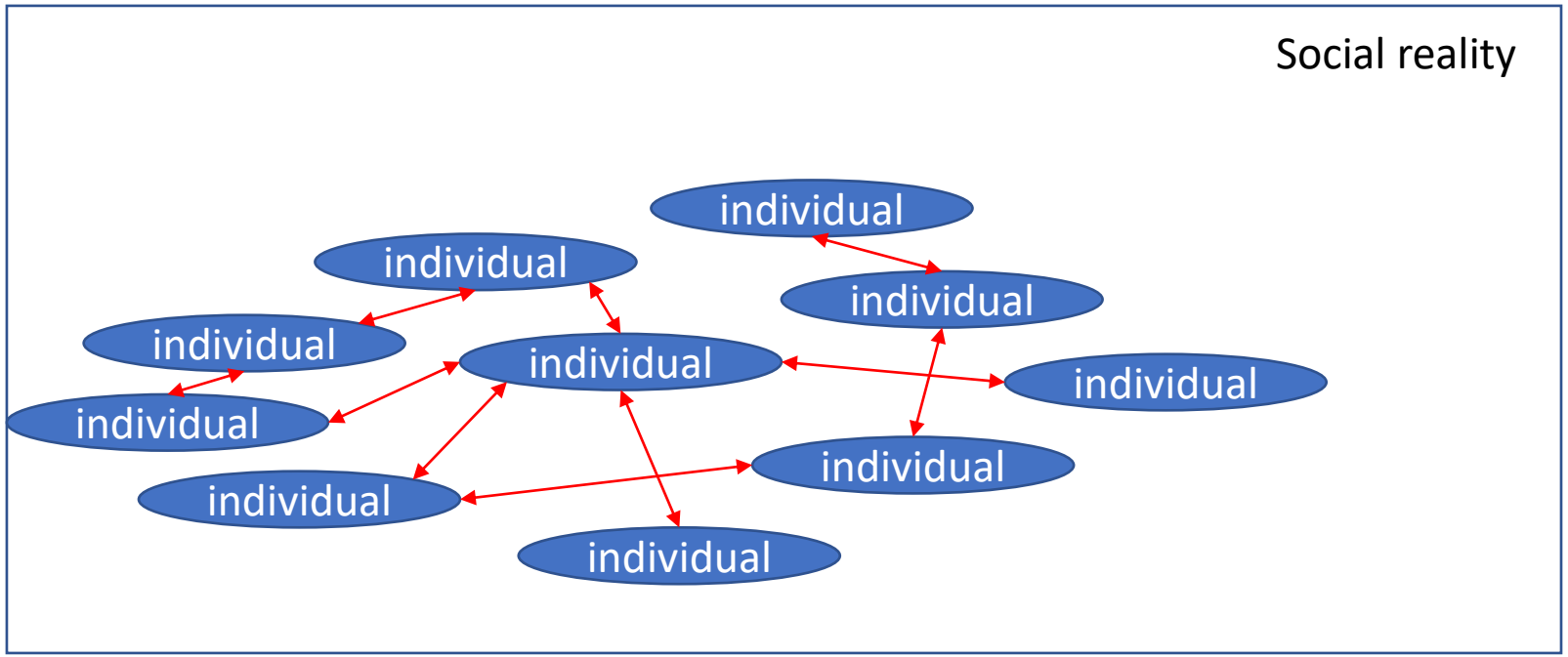
Today

- Intro and Motivation
- Formal Stuff
- Get-to-know-you
- Content Analysis & NLP
- Short break
- A very first look at quanteda

Session 1.1: Motivation

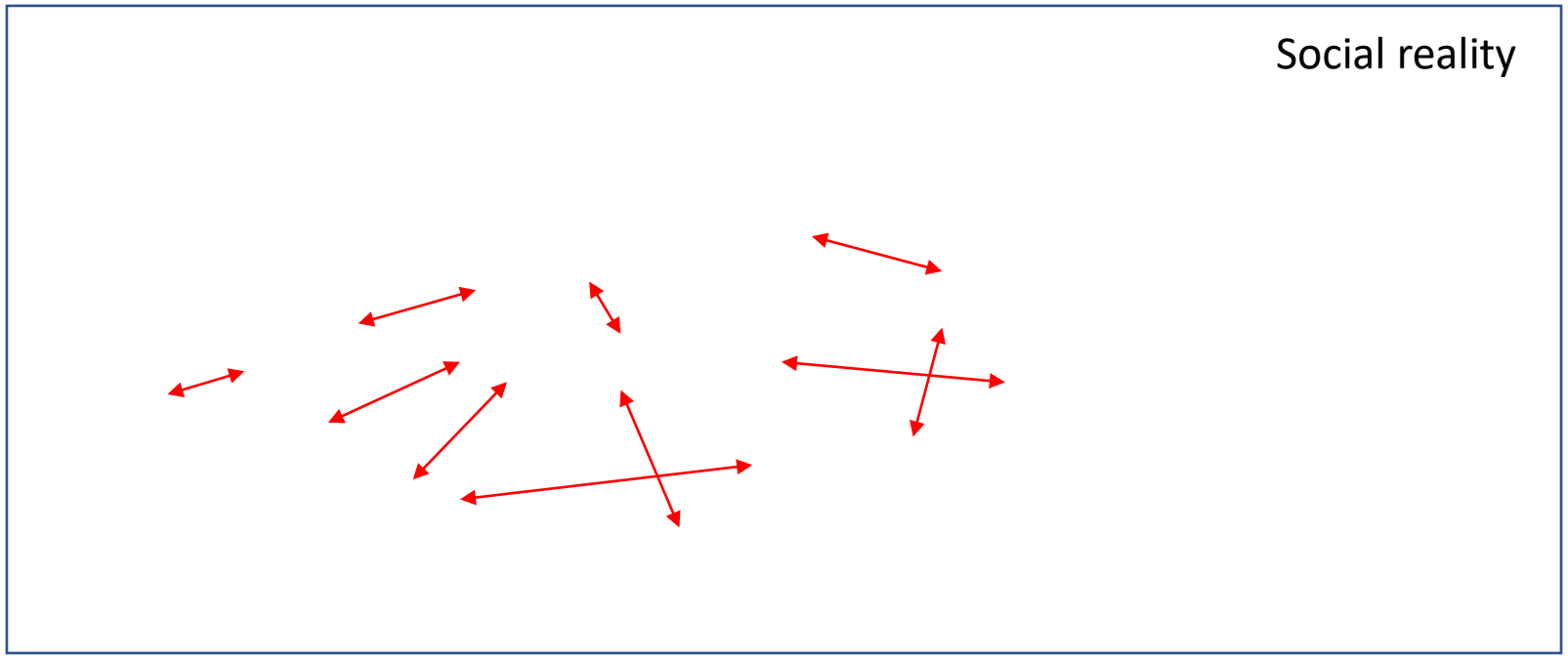
Text is important

Expression of Communication



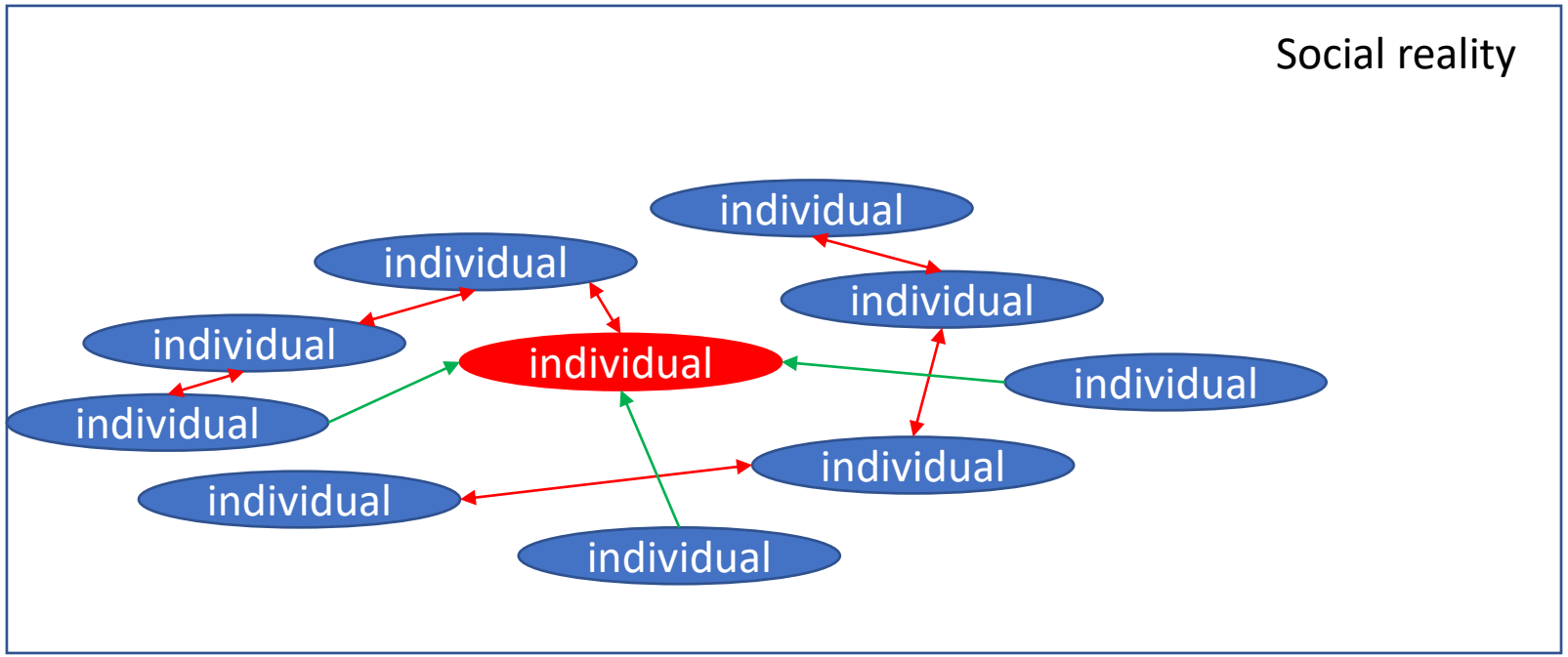
Text is important

Expression of Communication itself



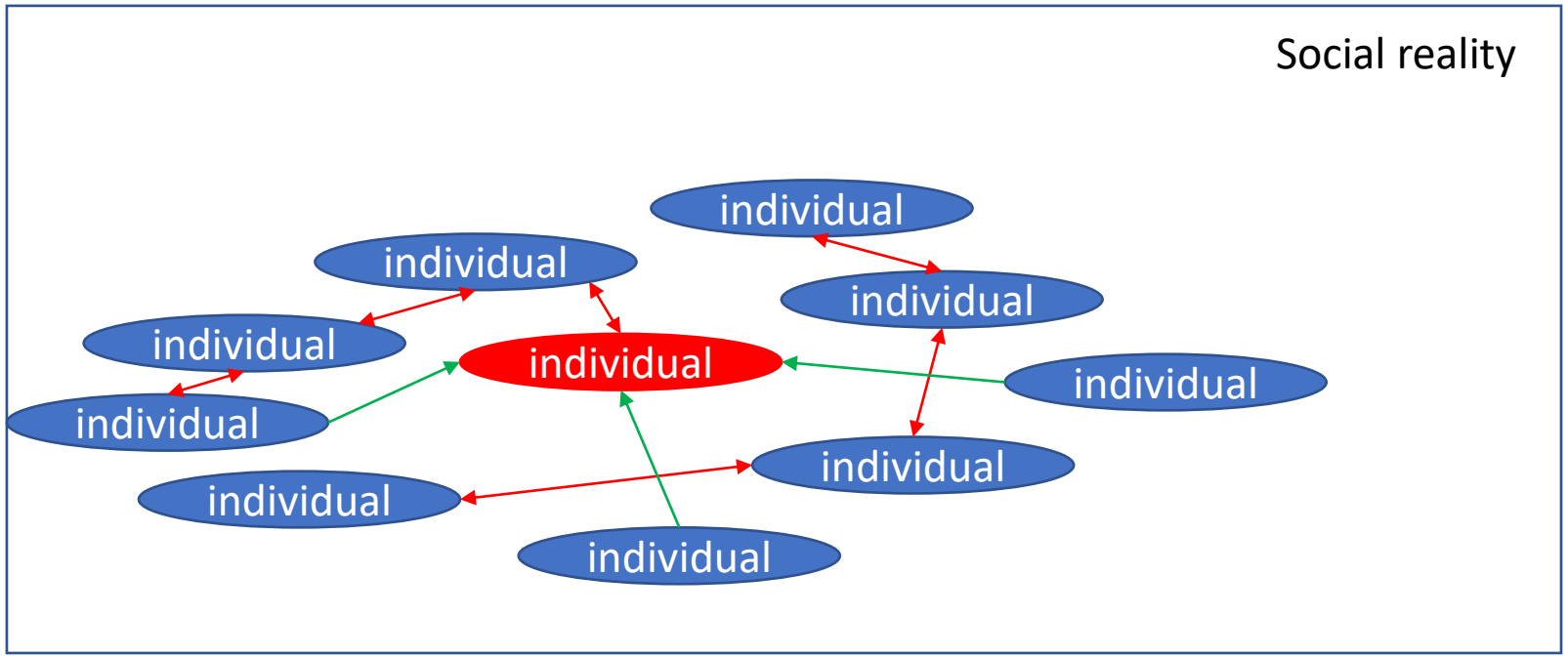
Text is important

Expression of perception: traits



Text is important

Expression of perception: traits



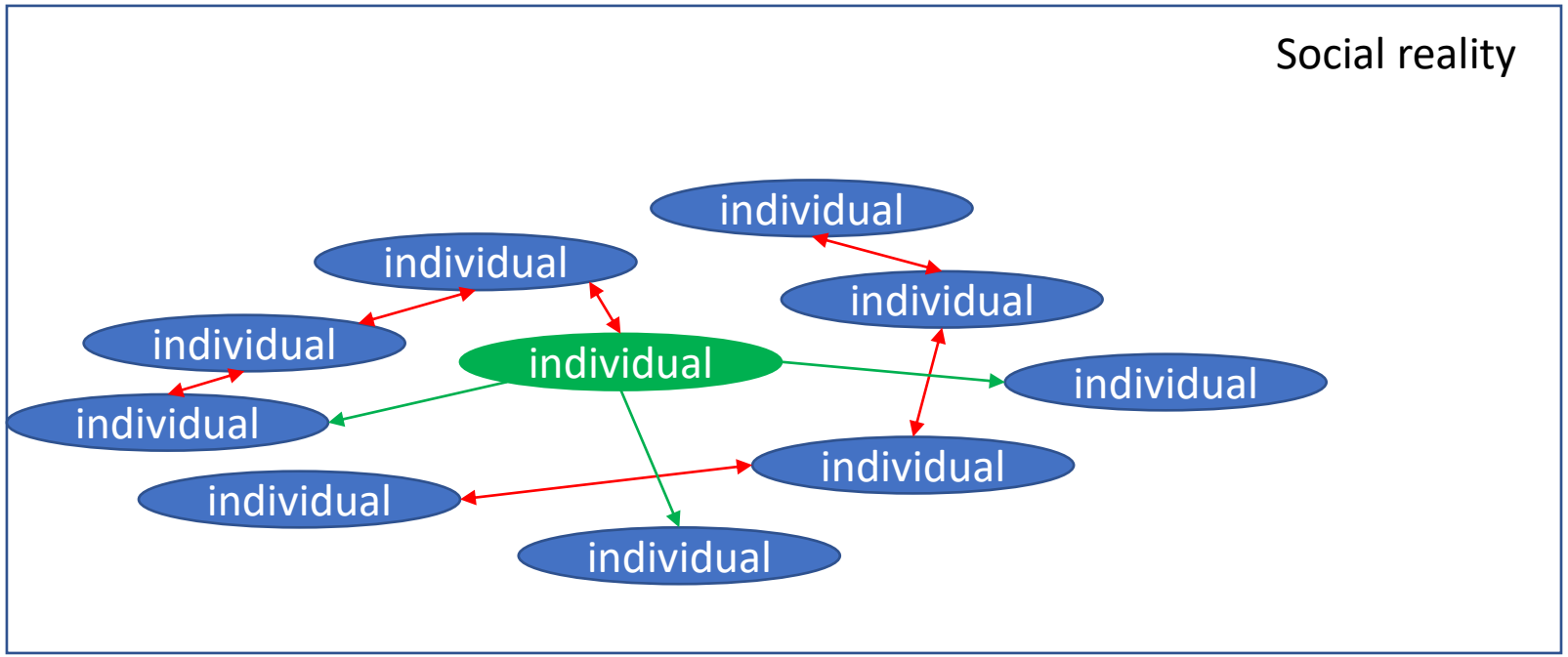
Populism

Framing

Agenda Setting

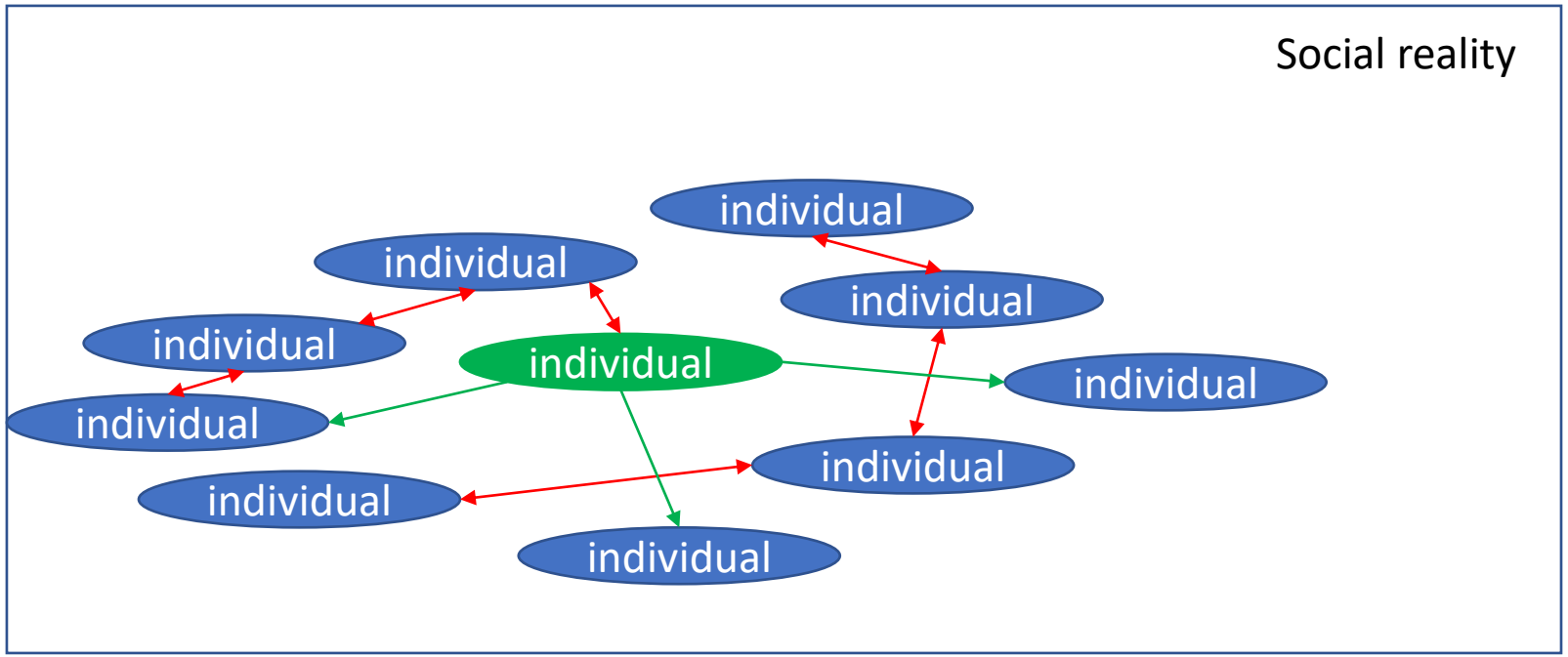
Text is important

Expression of strategy: infer intention



Text is important

Expression of strategy: infer intention



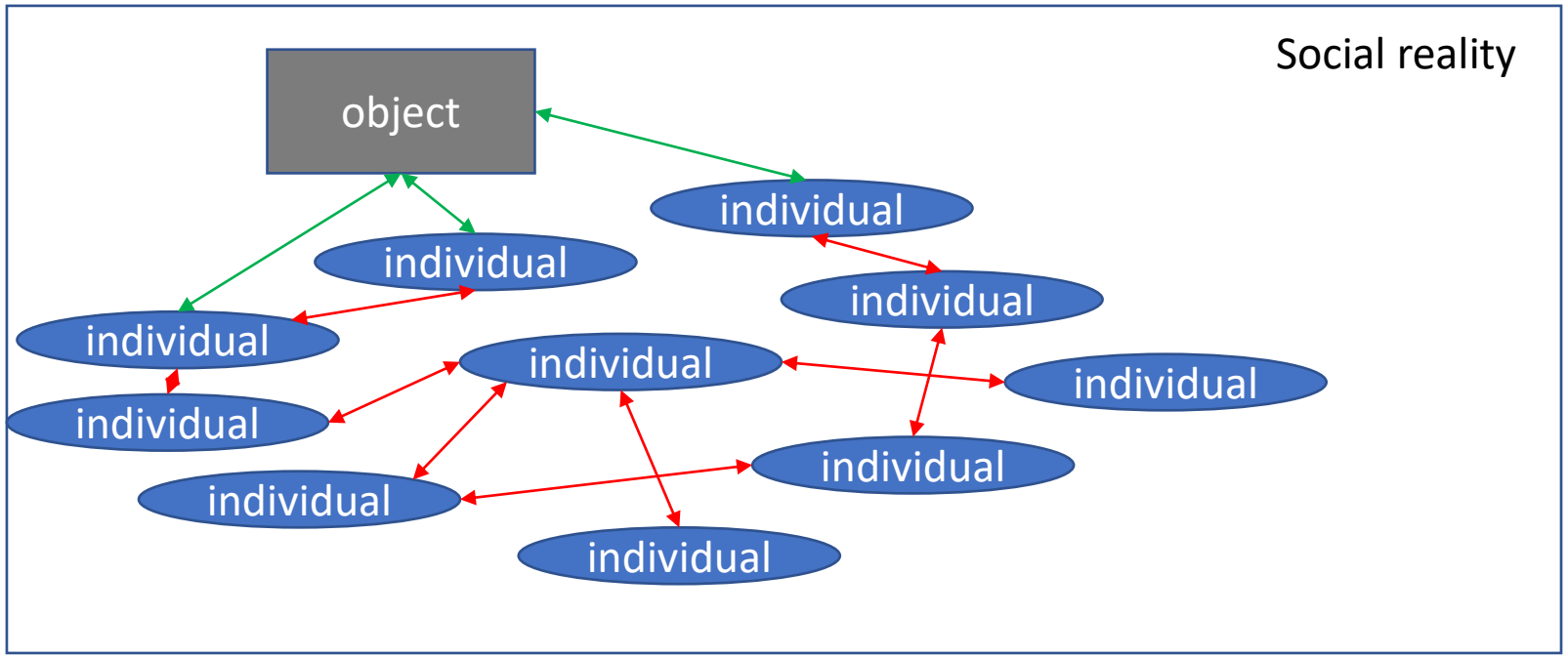
Populism

Framing

Agenda Setting

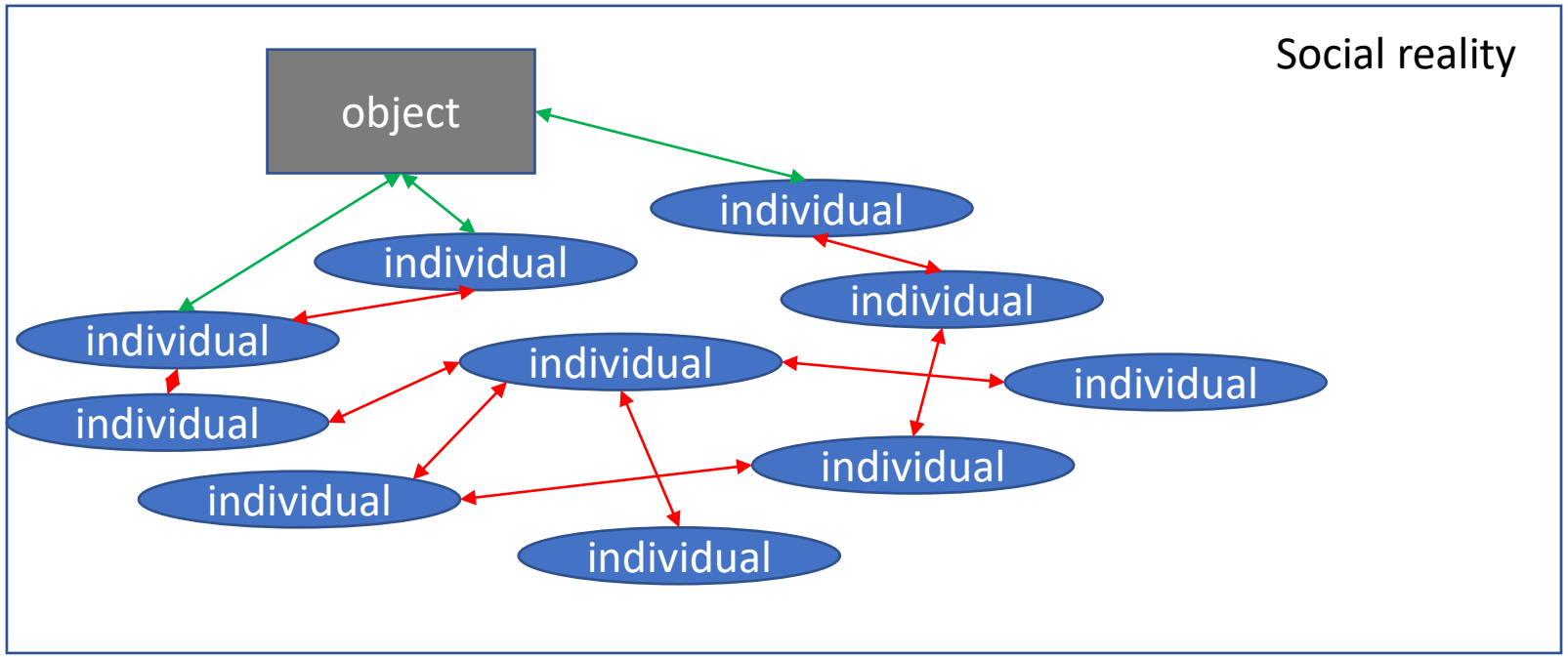
Text is important

Expression of evaluation: infer opinion



Text is important

Expression of evaluation: infer opinion



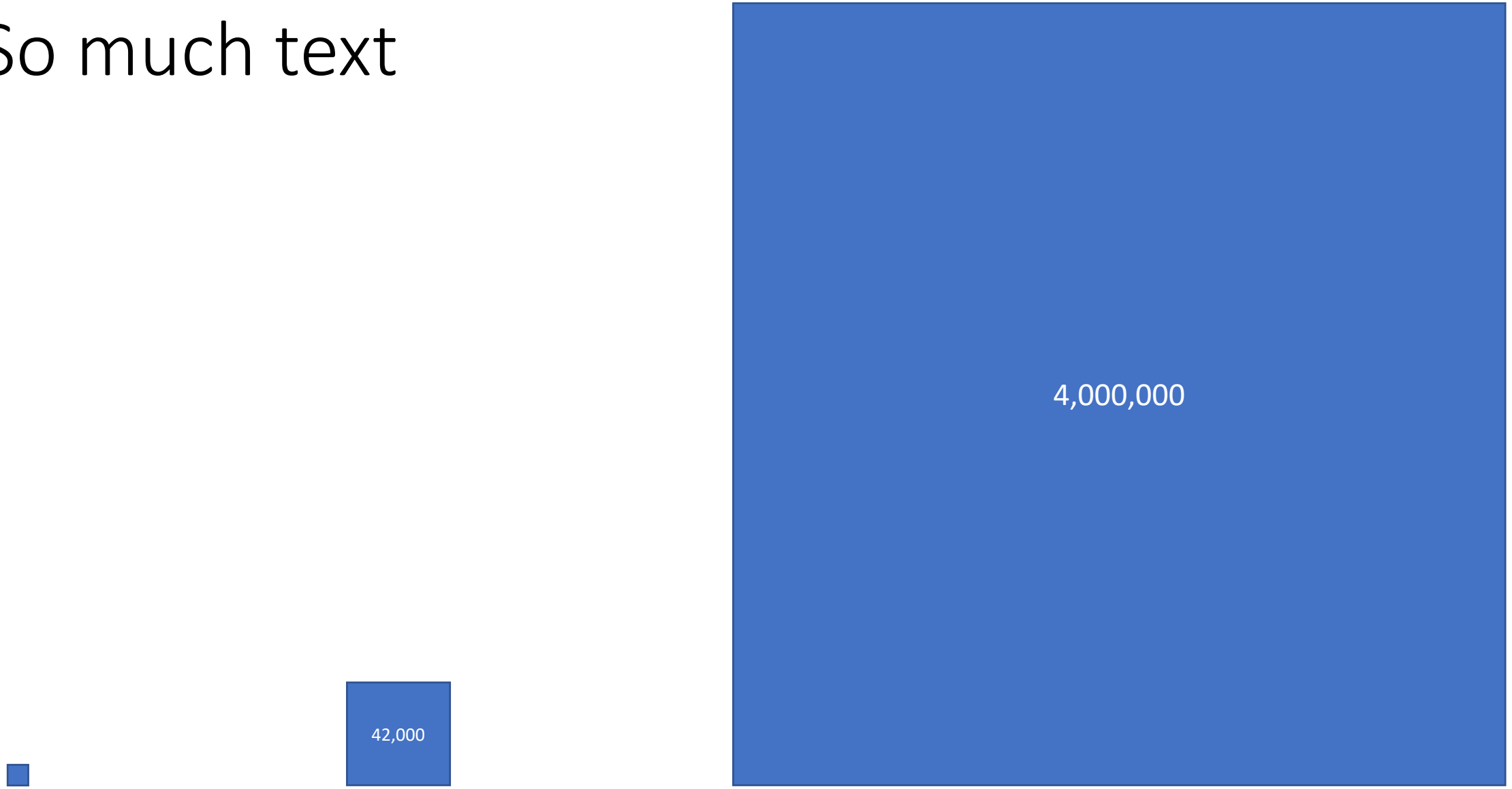
Sentiment analysis on products

Hate speech against politicians

What is Computational Text Analysis?

- Content Analysis
  - Inferring concepts relevant to research questions
  - From text

So much text



N of a survey

N of a survey

Sentences in 10 years of manifestos

Social media posts by candidates during the german Federal election 2021

Course Structure

- Introduction
- Text as Data
- Unsupervised Methods
- Supervised Learning
- Text in the Wild
- Your projects
- My solutions

Evaluation

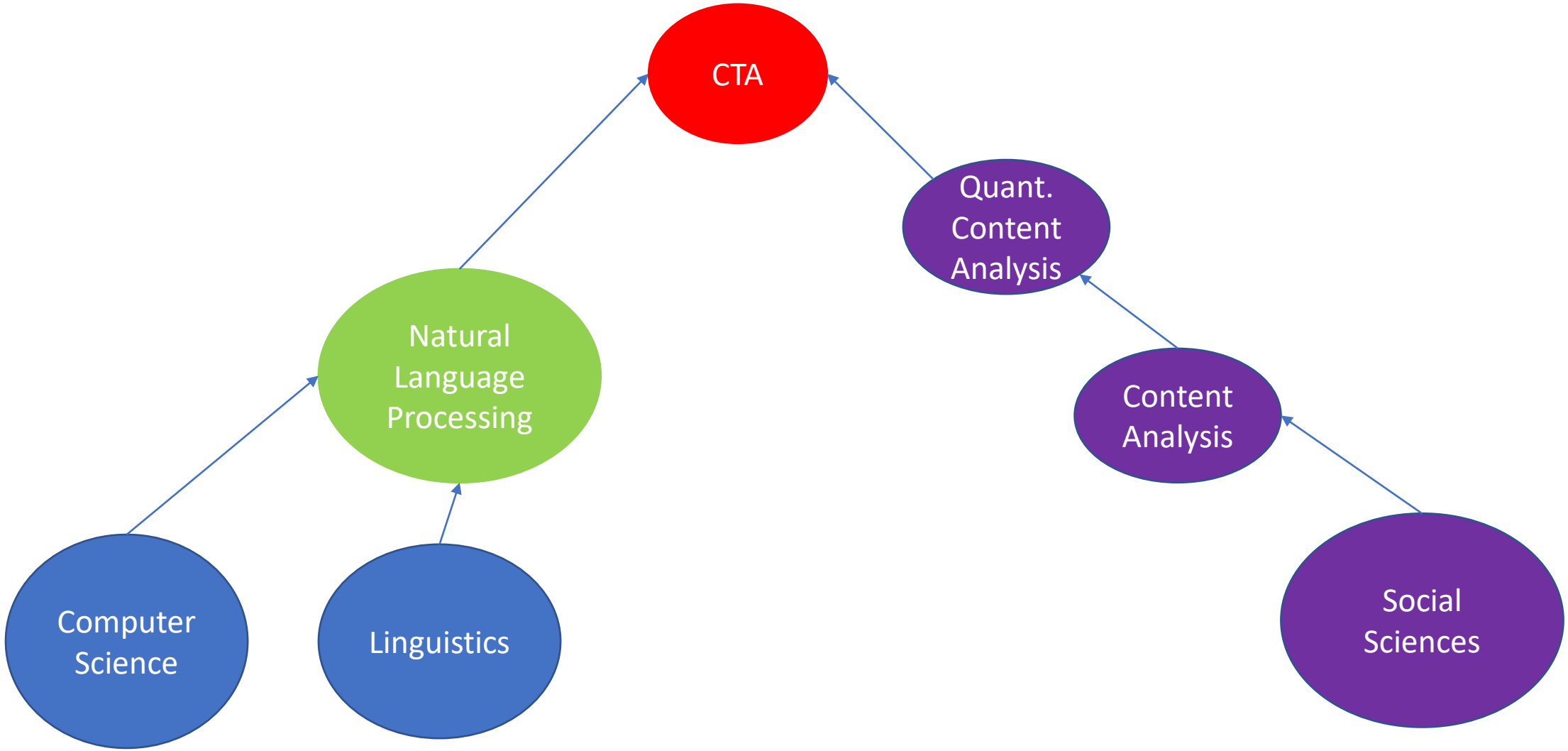
- You are grad students:
  - I don‘t care about grades
  - I want to help with YOUR research
- Learning by Doing:
  - Best way to learn: apply to your problem
  - Bring your own corpus!
  - Homework will always be: apply learned stuff to this corpus
- Your research is your term paper
  - Reviewer‘s perspective
  - Method focus

Who are you?

- What is your
  - Name
  - Research Area
  - Plan to do with Text analysis?

Session 1.2: Content Analysis

What is Computational Text Analysis?



Content Analysis

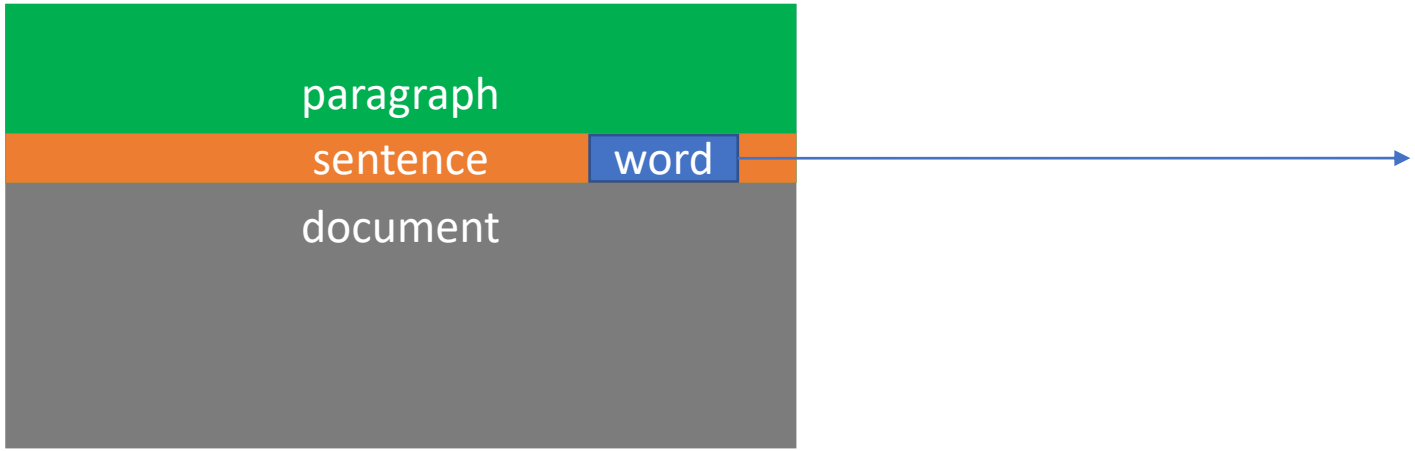
„Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use." (Krippendorf 2009)

- Qualitative
  - Discourse Analysis
  - Social Constructivist
- Quantitative
  - Manual (En-) Coding
  - Result: texts - categories

What can we do with it?

- Content that resides in Text
  - Information Extraction (NER)
  - Summarization
  - Text characteristics (readbility, sentiment)
- Inferences about the context/author/meaning
  - Inference of charateristics (Ideology, Saliency, Traits)
  - Inferences about intent (Strategies, Framing)

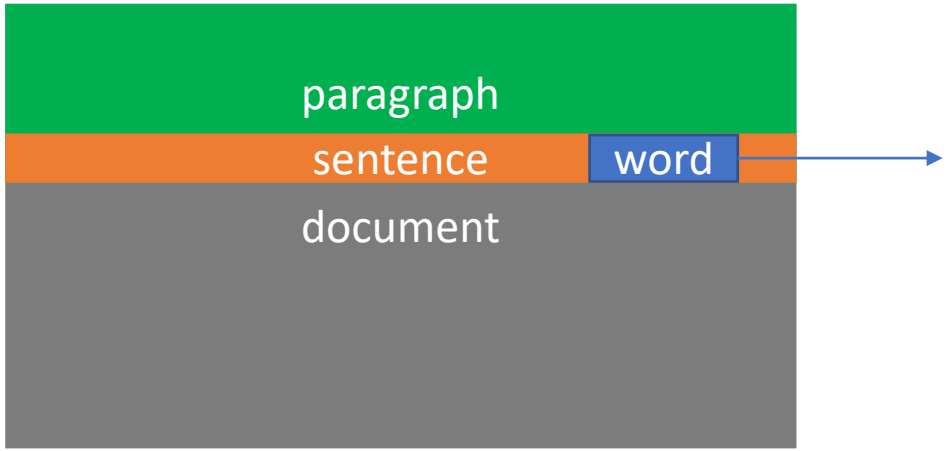
What to do with text?



How often occurs a term:

Linguistics
Targets (named entities)

What to do with text?



human

concept

„populism“

„sentiment“

„saliency“

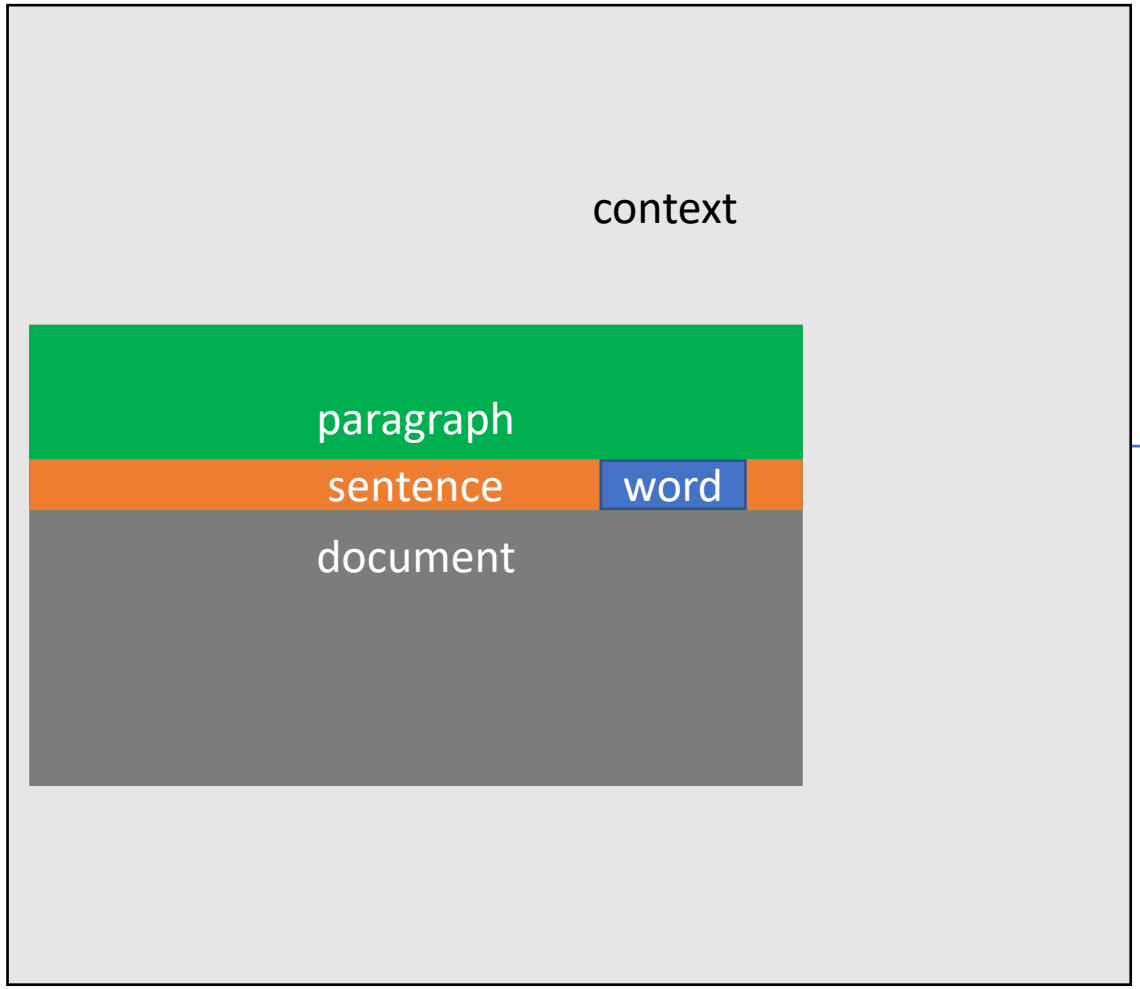
„ideology“

„orientation“

„personality“

„authorship“

What to do with text?



human

Analytical construct

concept

„populism“

„sentiment“

„saliency“

„ideology“

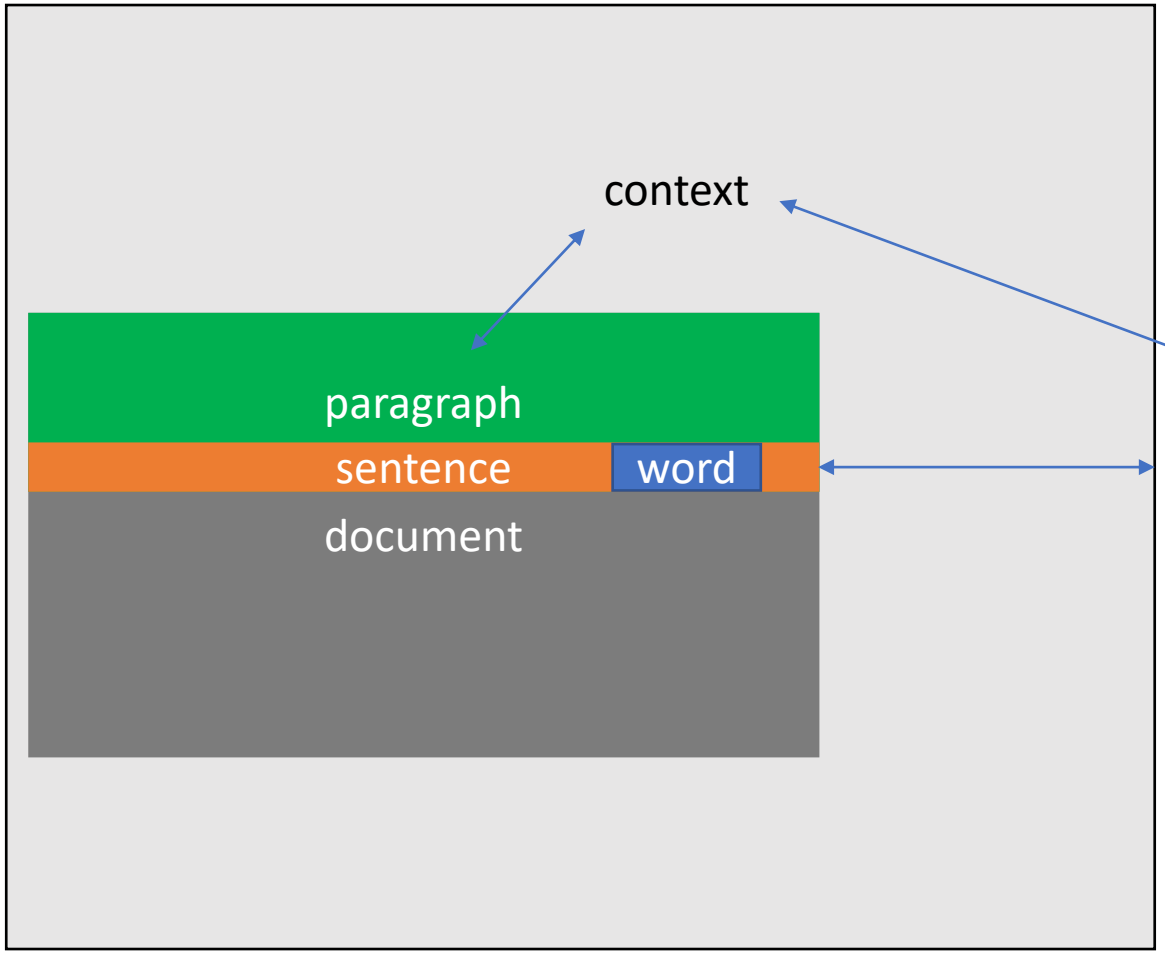
„orientation“

„personality“

„authorship“

“operationalize what the content analyst knows about the context […] how available texts are connected to the possible answers to the analyst's questions” Krippendorff 2009 , p 34-35

Automating content analysis



Analytical construct

concept

„populism“

„sentiment“

„saliency“

„ideology“

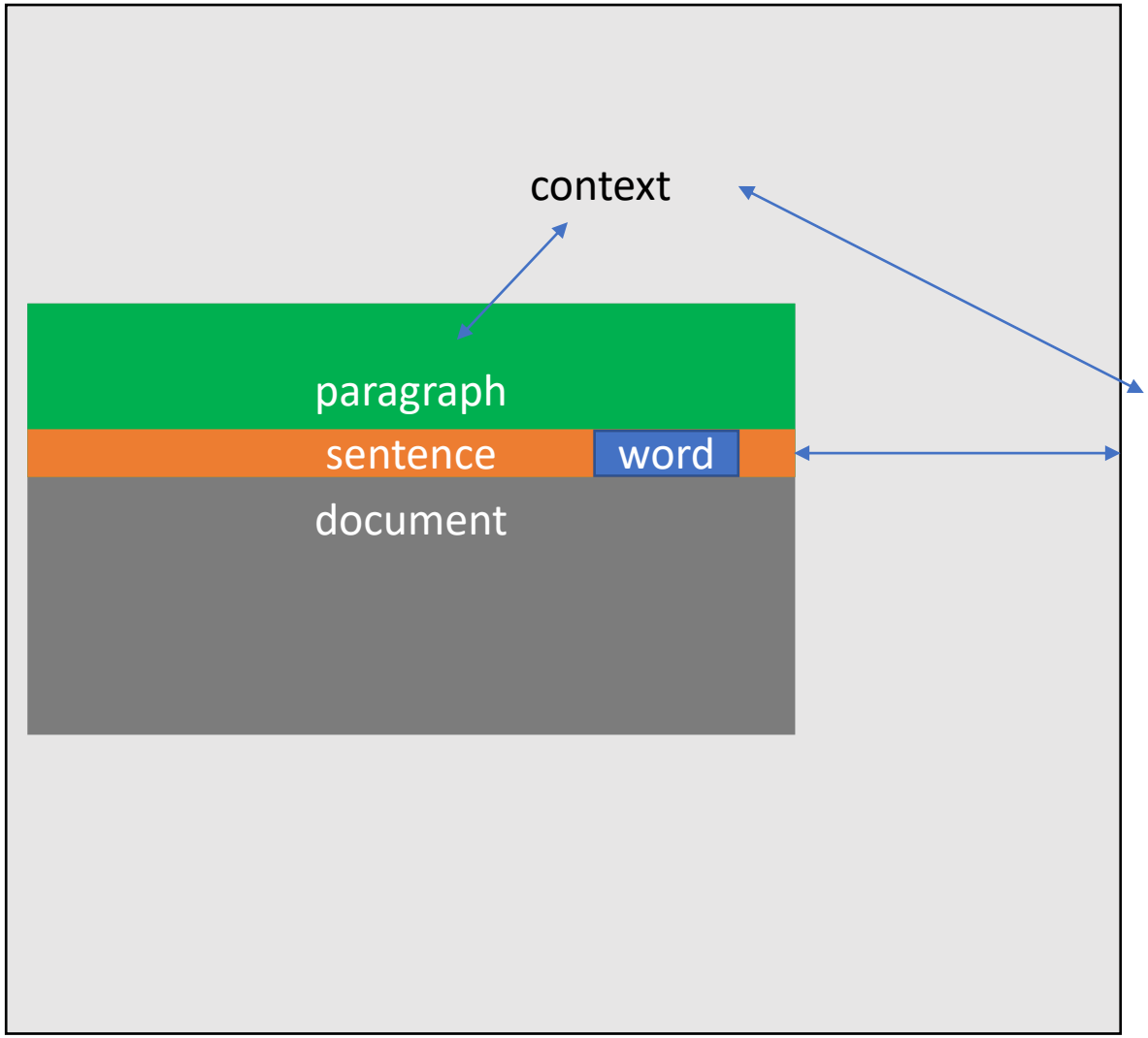
„orientation“

„personality“

„authorship“

A set of rules that maps words as used in this context to a concept

Replacing the coder



dictionary

„bad“
„horrendous“
„good“
„nice“

concept

„good movie“

Summary

- Computational Content Analysis:
  - a set of rules
  - That can map a concept of scientific interest
  - To the occurrence of words in context

Now: How do we build these rules so a computer can apply them to text?

Session 1.3: Natural Language Processing

Natural Language Processing

- The technique how computers understand „natural language“
- Using:
  - Linguistics
  - Computer Science

Using rules of grammar and computation to „understand“ the meaning of text.

The Problem of Natural Language

- Complex (many ways to say the same)
  - Synonyms
  - Sentence Structure
  - Spelling Errors
  - Metaphors
- Context dependent (many meanings of the same)
  - Homonyms
  - Irony
  - Implicit meaning

History

- Corpus Linguistics
  - Annotation of Corpora by Hand
  - Rule – based analysis
- Increasing demand: Internet
- Moore‘s Law
  - Statistical Learning
  - Neural Networks

Rule-Based

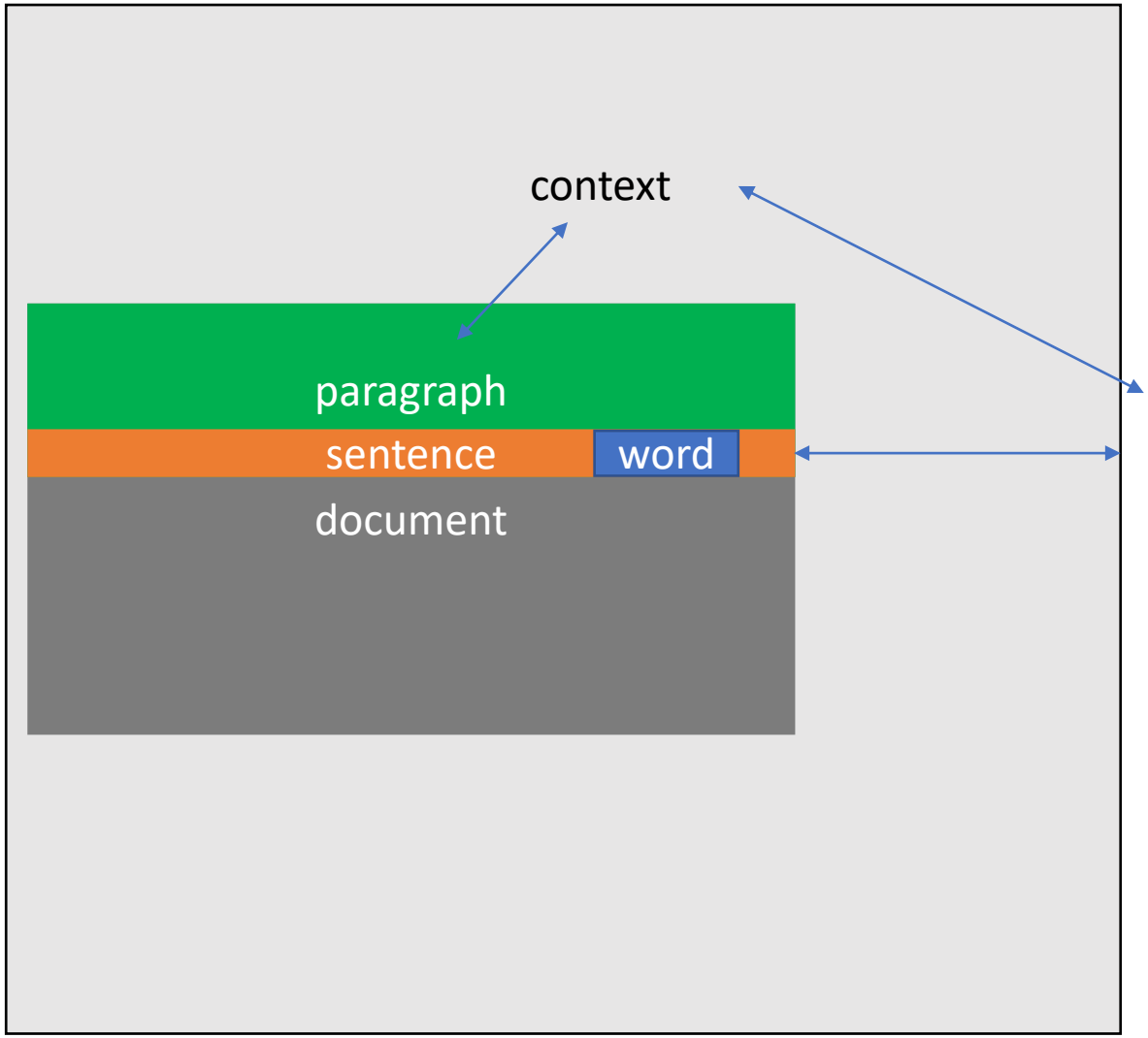
Preprocessing

- Words end with a space
- Sentences end with punctuation
- Applying the rules of using language
  - Lemmatization
  - Stemming

Concepts

- Dictionary approaches
- If „bad“, negative

Mapping to words



dictionary

„bad“
„horrendous“
„good“
„nice“

concept

„good movie“

It's not a laugh out loud comedy - it's not supposed to be!rebeccajmoore-971634 January 2022
This was pure satire and frightening true to life with the social media obsessed generation who would rather spend all day hearing about the latest pop star scandal than the possibility the world is about to end. Clever, fast moving, crazy at times but the message is clear - if the world were about to end the majority would probably get their finest designer gear on to video themselves waiting ! 10/10

Satire for the ages 1 January 2022
If you love a good satire, then look no further. If you have more than five brain cells, look for something a little bit more intelligent. An all star cast that act their parts beautifully. Good effects and cinematography. It's just too obvious. 4/10

Boring and stooped 7 January 2022
Jonah hill was such a idiot and the president, the president was even more stoopid The whole cast was on some stoopid ish..except DiCaprio. Definitely don't recommend this movie to anyone. 1/10

Boring, slow, waste of timeartera-273984 January 2022
Honestly, if this is the level of Hollywood humor, I do not see the humor, but although I continue to give Netflix opportunities, they do not disappoint, one day they will make a 5/10 movie. 1/10

It's not a laugh out loud comedy - it's not supposed to be!rebeccajmoore-971634 January 2022
This was pure satire and frightening true to life with the social media obsessed generation who would rather spend all day hearing about the latest pop star scandal than the possibility the world is about to end. Clever, fast moving, crazy at times but the message is clear - if the world were about to end the majority would probably get their finest designer gear on to video themselves waiting ! 10/10

Satire for the ages 1 January 2022
If you love a good satire, then look no further. If you have more than five brain cells, look for something a little bit more intelligent. An all star cast that act their parts beautifully. Good effects and cinematography. It's just too obvious. 4/10

Boring and stooped 7 January 2022
Jonah hill was such a idiot and the president, the president was even more stoopid The whole cast was on some stoopid ish..except DiCaprio. Definitely don't recommend this movie to anyone. 1/10

Boring, slow, waste of timeartera-273984 January 2022
Honestly, if this is the level of Hollywood humor, I do not see the humor, but although I continue to give Netflix opportunities, they do not disappoint, one day they will make a 5/10 movie. 1/10

It's not a laugh out loud comedy - it's not supposed to be!rebeccajmoore-971634 January 2022
This was pure satire and frightening true to life with the social media obsessed generation who would rather spend all day hearing about the latest pop star scandal than the possibility the world is about to end. Clever, fast moving, crazy at times but the message is clear - if the world were about to end the majority would probably get their finest designer gear on to video themselves waiting ! 10/10

Satire for the ages 1 January 2022
If you love a good satire, then look no further. If you have more than five brain cells, look for something a little bit more intelligent. An all star cast that act their parts beautifully. Good effects and cinematography. It's just too obvious. 4/10

Boring and stooped 7 January 2022
Jonah hill was such a idiot and the president, the president was even more stoopid The whole cast was on some stoopid ish..except DiCaprio. Definitely don't recommend this movie to anyone. 1/10

Boring, slow, waste of timeartera-273984 January 2022
Honestly, if this is the level of Hollywood humor, I do not see the humor, but although I continue to give Netflix opportunities, they do not disappoint, one day they will make a 5/10 movie. 1/10

It's not a laugh out loud comedy - it's not supposed to be!rebeccajmoore-971634 January 2022
This was pure satire and frightening true to life with the social media obsessed generation who would rather spend all day hearing about the latest pop star scandal than the possibility the world is about to end. Clever, fast moving, crazy at times but the message is clear - if the world were about to end the majority would probably get their finest designer gear on to video themselves waiting ! 10/10

Satire for the ages 1 January 2022
If you love a good satire, then look no further. If you have more than five brain cells, look for something a little bit more intelligent. An all star cast that act their parts beautifully. Good effects and cinematography. It's just too obvious. 4/10

Boring and stooped 7 January 2022
Jonah hill was such a idiot and the president, the president was even more stoopid The whole cast was on some stoopid ish..except DiCaprio. Definitely don't recommend this movie to anyone. 1/10

Boring, slow, waste of timeartera-273984 January 2022
Honestly, if this is the level of Hollywood humor, I do not see the humor, but although I continue to give Netflix opportunities, they do not disappoint, one day they will make a 5/10 movie. 1/10

Please open Script 1

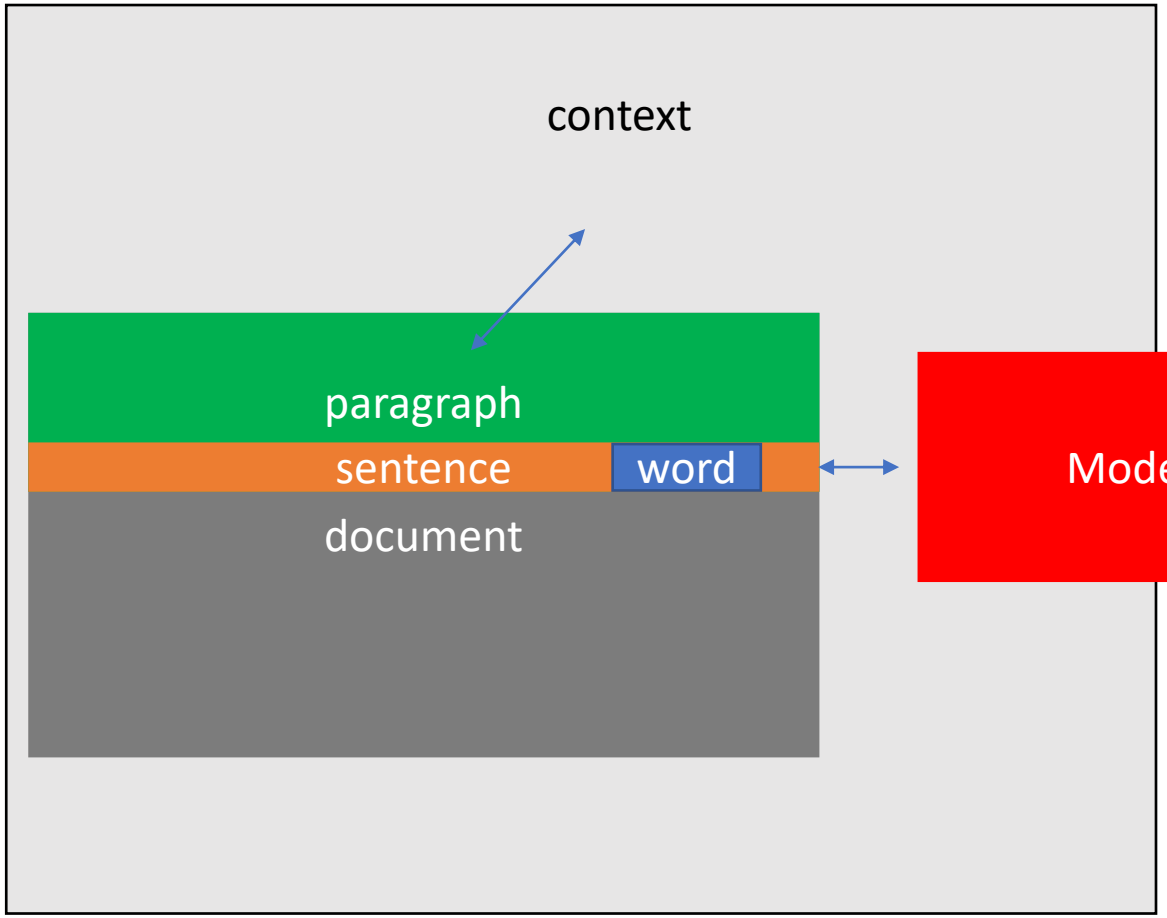
NLP for Content Analysis

- Plug in tasks from Text Mining:
  - Classification
  - Clustering
  - Information Extraction

- Plug in methods from NLP
  - Pattern Matching
  - Tokenization
  - Word Embeddings

- To identify analyitcal constructs relevant for social science

Your Research



Model

Incorporating context

concept

„your concept“